Detection of abnormalities and electricity theft using genetic Support Vector Machines

Conference Paper · December 2008

DOI: 10.1109/TENCON.2008.4766403 · Source: IEEE Xplore

5 authors, including:

Jawad Nagi
New York University
46 PUBLICATIONS **1,385** CITATIONS

SEE PROFILE

Syed Khaleel Ahmed
Universiti Tenaga Nasional (UNITEN)
47 PUBLICATIONS **973** CITATIONS

SEE PROFILE

Detection of Abnormalities and Electricity Theft using Genetic Support Vector Machines

J. Nagi, K. S. Yap, S. K. Tiong, *Member, IEEE*, S. K. Ahmed, *Member, IEEE*, A. M. Mohammad

***Abstract*—Efficient methods for detecting electricity fraud has been an active research area in recent years. This paper presents a hybrid approach towards Non-Technical Loss (NTL) analysis for electric utilities using Genetic Algorithm (GA) and Support Vector Machine (SVM). The main motivation of this study is to assist Tenaga Nasional Berhad (TNB) in Malaysia to reduce its NTLs in the distribution sector. This hybrid GA-SVM model preselects suspected customers to be inspected onsite for fraud based on abnormal consumption behavior. The proposed approach uses customer load profile information to expose abnormal behavior that is known to be highly correlated with NTL activities. GA provides an increased convergence and globally optimized SVM hyper-parameters using a combination of random and prepopulated genomes. The result of the fraud detection model yields classified classes that are used to shortlist potential fraud suspects for onsite inspection. Simulation results prove the proposed method is more effective compared to the current actions taken by TNB in order to reduce NTL activities.**

***Index Terms*—Support vector machine, genetic algorithm, electricity theft, non-technical loss, load profile.**

I. Introduction

ELECTRIC utilities lose large amounts of money each year due to fraud by electricity consumers. Electricity fraud can be defined as a dishonest or illegal use of electricity equipment or service with the intention to avoid billing charge. It is difficult to distinguish between honest and fraudulent customers. Realistically, electric utilities will never be able to eliminate fraud, however, it is possible to take measures to detect, prevent and reduce fraud [1].

Investigations are undertaken by electric utility companies to assess the impact of technical losses in generation, transmission and distribution networks, and the overall performance of power networks [2-3]. Non-technical losses (NTL) comprise one of the most important concerns for electricity distribution utilities worldwide. In 2004, Tenaga Nasional Berhad (TNB) the sole electricity provider in Peninsular Malaysia recorded revenue losses as high as USD 229 million a year as a result of electricity theft, billing errors and faulty metering [4].

In recent years, several data mining and research studies on fraud identification and prediction techniques have been carried out in the electricity distribution sector. These include Statistical Methods [5], Decision Trees [6], Artificial Neural Networks (ANNs) [7], Knowledge Discovery in Databases (KDD) [8], and Multiple Classifiers using cross identification and voting scheme [1]. Among these, load profiling is one of the most widely used [9], which is defined as the pattern of electricity demand of a customer over a period of time.

TNB in Malaysia is currently focusing on reducing its NTLs, which are estimated around 20% throughout Peninsular Malaysia. At present, customer installation inspections by TNB Distribution (TNBD) Division are carried out without any specific focus due to unavailability of a system for short listing possible fraud suspects. The approach proposed in this paper provides an intelligent system for assisting TNB inspection teams to increase effectiveness of their operation in reducing NTLs, and detecting fraudulent consumers based on load profiles of customers derived from the customer database. This system will increase fraud detection hit-rate for onsite inspection and reduce operational costs due to onsite inspection in monitoring NTL activities.

This paper presents a novel framework to detect NTL activities i.e., customers with abnormal consumption patterns indicating fraudulent activities. An automatic feature extraction method for load profiles with a combination of Support Vector Machines (SVMs) is used to identify fraud customers. This study uses historical customer data collected from TNBD. Customer consumption patterns are extracted using data mining techniques, which represent customer load profiles. Based on the assumption that load profiles contain abnormalities when a fraud event occurs, SVM classifies load profiles of customers for detection of fraud suspects. There are several different types of fraud that can occur, but our research concentrates only on scenarios where abrupt changes appear in load profiles, indicating fraudulent activities.

II. Genetic Algorithm

Genetic algorithm (GA) was first proposed by J. H. Holland in the 1960s, which applies the principles of evolution found in nature to find an optimal solution to an optimization problem [10].

In GA a solution is represented by a chromosome and the GA keeps a set (population) of chromosomes. Each element of a chromosome is referred to as a gene. The population evolves through a number of generations. Each generation is performed as follows. First, two solutions are selected in the population based on a certain probability distribution; they are

called parent chromosomes. Then, the crossover operation produces an offspring chromosome by combining the parents. The mutation operation then modifies the offspring chromosome with a low probability. The offspring can be locally improved by any other algorithm or heuristic. A generation is completed by replacing one of the members in the population with the offspring. A considerable number of generations are run until a user-defined convergence criterion is reached. Finally, the GA returns the optimized parameters or variables in the population as the solution [11].

III. Support Vector Machine

Support vector machines (SVMs) were introduced by Vapnik [12] in the late 1960s on the foundation of statistical learning theory. SVMs are a set a novel machine learning methods for classification and regression. In SVM, training is performed in a way such to obtain a quadratic programming (QP) problem. The solution to this QP problem is global and unique. For empirical data $(x_1, y_1),\ldots,(x_m, y_m) \in R^n \times \{-1,+1\}$ that are mapped by $\varphi: R^n \rightarrow F$ into a "feature space", the linear hyperplanes that divide them into two labeled classes can be mathematically represented as:

$$w \times \phi(x) + b = 0 \qquad w \in \Re^n, \quad b \in \Re \tag{1}$$

To construct an optimal hyperplane with maximum-margin and bounded error in the training data (soft margin), the following QP problem is to be solved:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$$
$$y_i\left(w \times \phi(x) + b\right) \geq 1 - \xi_i, \quad i = 1,2,...,m \tag{2}$$

The first term in cost function (2) makes maximum margin of separation between classes, and the second term provides an upper bound for the error in the training data. The constant $C \in [0,\infty)$ creates a tradeoff between the number of misclassified samples in the training set and separation of the rest samples with maximum margin. A way to solve (2) is via its Lagrange function. Given a kernel $K(x_i, y_i) = \varphi(x_i)\cdot\varphi(x_j)$, the Lagrange function of (2) is simplified to:

$$\max_{\alpha} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i, x_j) \tag{3}$$

$$w = \sum_{i=1}^{m} y_i\alpha_i\phi(x_i), \ \sum_{i=1}^{m}\alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad \forall i \tag{4}$$

From eq. (1) it is seen that the optimal hyperplane in the feature space can be written as the linear combination of training samples with $\alpha_i \neq 0$. These informative samples, known as *support vectors*, construct the decision function of the classifier based on the kernel function:

$$f(x) = \operatorname{sgn}\left(\sum_{i,=1}^{m} y_i\alpha_i k(x, x_j) + b\right) \tag{5}$$

Kernel functions in SVMs are selected based on the data structure and type of the boundaries between the classes. The representative and widely applied kernel function based on Euclidean distance is the radial basis function (RBF) kernel, also known as the Gaussian kernel [13]:

$$K^{RBF}(x_i, x_j) = \exp\left(-\gamma\left\|x_i - x_j\right\|^2\right) \tag{6}$$

where $\gamma > 0$ is the RBF kernel parameter. The RBF kernel induces an infinite-dimensional kernel space, in which all image vectors have the same norm, and the kernel width parameter "γ" controls the scaling of the mapping [13]. This paper employs LIBSVM [14], a library for support vector machines, as the core SVM classifier and conducts multiclass classifications using the "*One Against One*" or OAO method.

IV. Methodology

The fraud detection system presented in this paper is developed as standalone GUI software in Microsoft Visual Basic 6.0 using LIBSVM v2.86 [14]. The computer used for testing is a Dell PowerEdge 840 workstation with Windows XP, a 2.40 GHz Intel Quad-core Xeon X3320 Processor with 4 GB of RAM. Time elapsed for obtaining detection results from the customer database is approximated to be 2.3 seconds per customer. The proposed framework for fraud electricity customer detection is shown in Fig. 1.

A. Data Acquisition

Electricity customer consumption data from TNBDs electronic-Customer Information Billing System (e-CIBS) was obtained for Kuala Lumpur Barat station. The e-CIBS data consisted of 265,870 customers for a period of 25 months i.e., from June 2006 to June 2008.

B. Customer Filtering and Selection

The raw e-CIBS data obtained from TNBD was filtered for extraction of customer load profiles and features. Hence, data mining techniques using database querying were applied for:

- Removing repeating customers in monthly data.
- Removing customers having no consumption (0 kWh) throughout the entire 25 month period.
- Removing customers who are not present within the entire 25 month data i.e., removing new customers registered after the first month.

After customer filtering only 186,968 customer records remained. The main tasks involved in preprocessing the raw e-CIBS data for SVM classification are illustrated in Fig. 2.

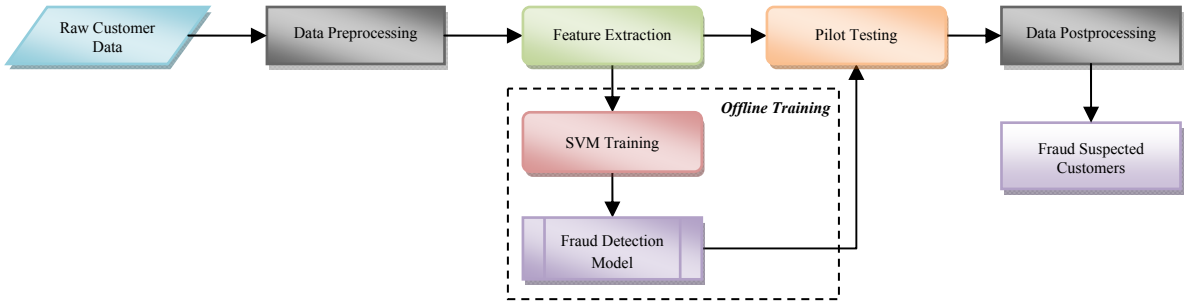

Fig. 1. Proposed framework for fraud electricity customer detection

C. Feature Extraction

From the 25 month customer database, 24 daily average consumption values were extracted for each customer, corresponding to features. These features relate to customer load profiles. For a selected group of M customers, each load profile is characterized by a vector $x^{(m)} = \{x_h^{(m)},\ h = 1,\ldots,H\}$, where $H = 24$ corresponds to time domain intervals based on average daily kWh consumption features representing the load profile. Therefore, the whole set of data is $X = \{x^{(m)},\ m = 1,\ldots,M\}$. Daily average consumption features for each customer were extracted using:

$$x_h^{(m)} = \frac{P_{h+1}}{D_{h+1} - D_h},\quad h = 1,2,\ldots,24 \tag{7}$$

where P_{h+1} represents the monthly power consumption of the following month, and $D_h - D_{h+1}$ represents the difference of days with respect meter reading date between the following and current months.

CWR (Credit Worthiness Rating) data from the e-CIBS data was taken as an additional feature for the fraud detection model. CWR is automatically generated from TNBDs billing system and is targeted to identify customers intentionally avoiding to pay bills and delaying payments. In the customer database CWR value ranges from 0.00 to 5.00, where 0.00 represents minimum CWR and 5.00 represents maximum CWR. Since CWR changes frequently based on monthly payment status of customers, averaged CWR for each customer over a period of 25 months was represented as an additional feature. Therefore, 25 features were selected for the SVM classifier i.e., 24 daily average kWh consumption features and 1 CWR feature.

D. Data Preprocessing

Real world data sets tend to be noisy and inconsistent. Therefore, to overcome these problems, data mining techniques using statistical methods were implemented. Customer data with estimated monthly kWh consumptions (cases where meter readers are unable to record meter readings, due to customers not being present at their residence) were preprocessed and converted to normal consumption values, to smoothen out noise.

E. Data Normalization

The load data needs to be represented using a normalized scale for the SVM classifier. Therefore, the daily average kWh consumption feature data was normalized as follows:

$$NL = \frac{L - \min(L)}{\max(L) - \min(L)} \tag{8}$$

where L represents the current kWh consumption of the customer, and $min(L)$ and $max(L)$ represent the minimum and maximum values in the 24 month consumption feature set. Typical load profiles of customers were then established, with each load profile being represented by the 24 normalized daily average kWh consumption features.

F. Feature Adjustment

All 25 features were given a label, where the labels are represented by integer values. Normalized feature values with labels are represented as a LIBSVM feature file [14], denoted by the matrix W, in the form:

$$W = \begin{bmatrix} l_{11}:x_{11} & \cdots & l_{1k}:x_{1k} & \cdots & l_{1M}:x_{1M} \\ l_{21}:x_{21} & \cdots & l_{2k}:x_{2k} & \cdots & l_{2M}:x_{2M} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ l_{p1}:x_{p1} & \cdots & l_{pk}:x_{pk} & \cdots & l_{pM}:x_{p1M} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ l_{N1}:x_{N1} & \cdots & l_{Nk}:x_{Nk} & \cdots & l_{NM}:x_{NM} \end{bmatrix} \tag{9}$$

where l represents the feature label, x represents the normalized feature value, $M = 25$ is the total number of features and $N = 186{,}968$ is total number of customers in the feature file.

G. SVM Classification

Load profiles were classified into categories according to their typical behavior and atypical behavior content. In this study a 4-class SVM classifier is used to represent four different types of load profiles. The feature file in eq. (9) indicated that out of 186,968 customers, 1171 customers were previously inspected by TNBD to be fraud cases.

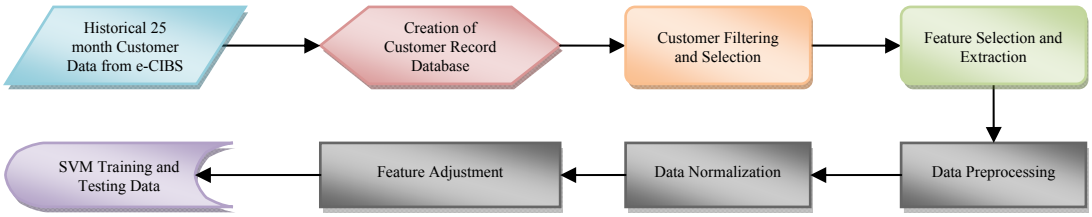

Fig. 2. Flowchart for preprocessing raw e-CIBS data for SVM classification

Manual inspection was done on each of the 1171 cases to identify load profiles in which abrupt changes appear clearly, (see Fig. 3) indicating *Confirm Fraud Suspects*. These cases were tagged as Class 1. Load profiles in which abrupt changes appeared, but oscillations are also present indicated *Unconfirmed Fraud Suspects,* and were tagged as Class 2. Similarly, inspection was done on a set of 1000 load profiles, not inspected by TNBD, to represent *Confirm Clean Suspects* (see Fig. 4) and *Unconfirmed Clean Suspects* i.e., Class 3 and Class 4 respectively. Out of the 1171 fraud cases in the customer database, only 131 cases were used in the classifier. The other 1040 load profiles did not have any abnormal patterns matching to fraud cases; i.e., these customers possibly committed electricity theft before the two year period for which there was no customer data.

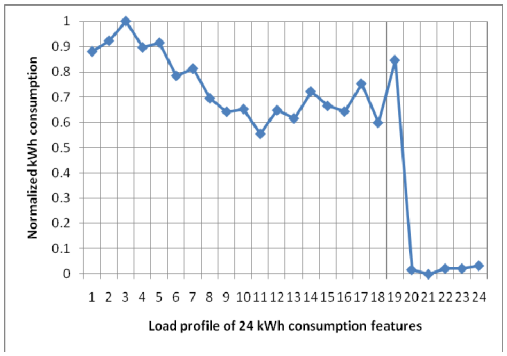

Fig. 3. Load profile of a typical fraud customer over a period of two years

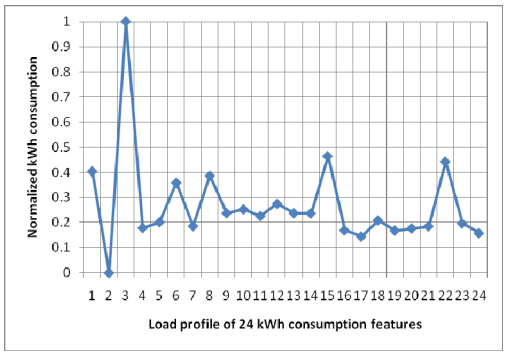

Fig. 4. Load profile of a clean customer over a period of two years

The classifier having unbalanced classes is weighted to balance the sample ratio. Weights are adjusted by calculating ratios for each class, by dividing the total number of classifier samples with individual class samples. Class weights are multiplied by a weight factor, w_f= 100 to achieve satisfactory weight ratios for the C-SVM training model. Specifications of the multi-class C-SVM classifier are illustrated in Table I.

TABLE I
SPECIFICATIONS OF MULTI-CLASS SVM CLASSIFIER

Class	Training samples	Support vectors	Class weightage
1	72	53	629.16
2	59	51	767.70
3	72	50	629.16
4	250	93	181.20

H. GA Optimization

A hybrid approach of GA-SVM is used to globally optimize SVM hyper-parameters for the following Dual Lagrangian Optimization (DLO) problem:

$$L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \qquad (10)$$

Since the SVM uses the RBF kernel, the hyper-parameters include: Lagrange multipliers (α_1, α_2,...,α_i), C and γ. In GA for optimization, all corresponding parameters are directly coded to form a chromosome. Consequently, the chromosome X is represented as X = {α_1, α_2,…,α_i, $p1$, $p2$}, where i represents the number of training features, $p1$ and $p2$ denote "C" the cost of error and "γ" is the RBF kernel parameter respectively as illustrated in Fig. 5.

α_1	α_2	α_3	α_4	…	C	γ

Fig. 5. Layout of GA chromosome

For the purpose of reducing the search space of GA and to achieve better genes a Prepopulated Database (PPD) is introduced. The PPD initially stores the 10 best chromosomes from the first training process. During subsequent training processes, the 10 best genomes are mixed with the random populated genomes. All chromosomes are evaluated by GA and the 10 fittest chromosomes are restored into the PPD.

The hybrid GA-SVM process is based on the survival principle of the fittest member in a population, which retains its genetic information by passing it on from generation to generation. The process of GA for SVM hyper-parameter optimization is described as follows:

1. *Initialization*: Generate a random initial population of n chromosomes (suitable solutions for the problem).
2. *Fitness Evaluation*: Evaluate the fitness $f(x)$ of each chromosome x in the problem. In this problem, the DLO function in eq. (10) is used as the fitness function.
3. *Selection*: Select two parent chromosomes according to their fitness for reproduction using the Roulette Wheel method. The area of the slice is proportional to the chromosome fitness ratio R_f and is calculated using:

$$R_f = \frac{f(i)}{\sum_{i=1}^{n} f(i)} x100\% \qquad (11)$$

where $f(i)$ is the fitness of the ith chromosome.

4. *Crossover*: Form new offspring (children) from the parents using a single-point crossover probability.
5. *Mutation*: Mutate the new offspring at each position using a uniform mutation probability measure.
6. *Next Generation*: Form the PPD for the next generation.
7. *Test:* If the number of generations exceeds the initialized value, then stop and return the best chromosomes in the current population as the solution.
8. *Loop:* Go to step 2.

For each of the 10 pairs of (C, γ) obtained from the PPD, performance was measured by training 70% of the classifier data and testing the other 30%. For GA optimization, several parameters were tested. Experimentally it was found that GA parameters illustrated in Table II, were the most suitable parameters for obtaining the highest SVM cross-validation (CV) accuracy and fraud detection hit-rate.

TABLE II
GA PARAMETERS USED FOR SVM HYPER-PARAMETER OPTIMIZATION

GA parameter	Value
Maximum Generation	500
Population Size	1000
Crossover Rate	0.8
Mutation Rate	0.025

From all the 10 pairs of (C, γ), optimal hyper-parameters selected were: C = 20.4726 and γ = 0.2608. The hybrid GA-SVM optimization and training engine is illustrated in Fig. 6. Using this (C, γ) parameter set, the highest 10-fold CV accuracy achieved was 92.58%. The reason for using 10-fold CV is to ensure the model does not overfit the training data.

I. SVM Prediction and Pilot Testing

For detecting fraudulent electricity consumers the trained fraud detection model was tested with TNBD e-CIBS data from three towns in the state of Kelantan in Malaysia. These towns are listed in Table III. Results obtained from pilot testing carried out indicated an average LIBSVM prediction accuracy of 81.63%.

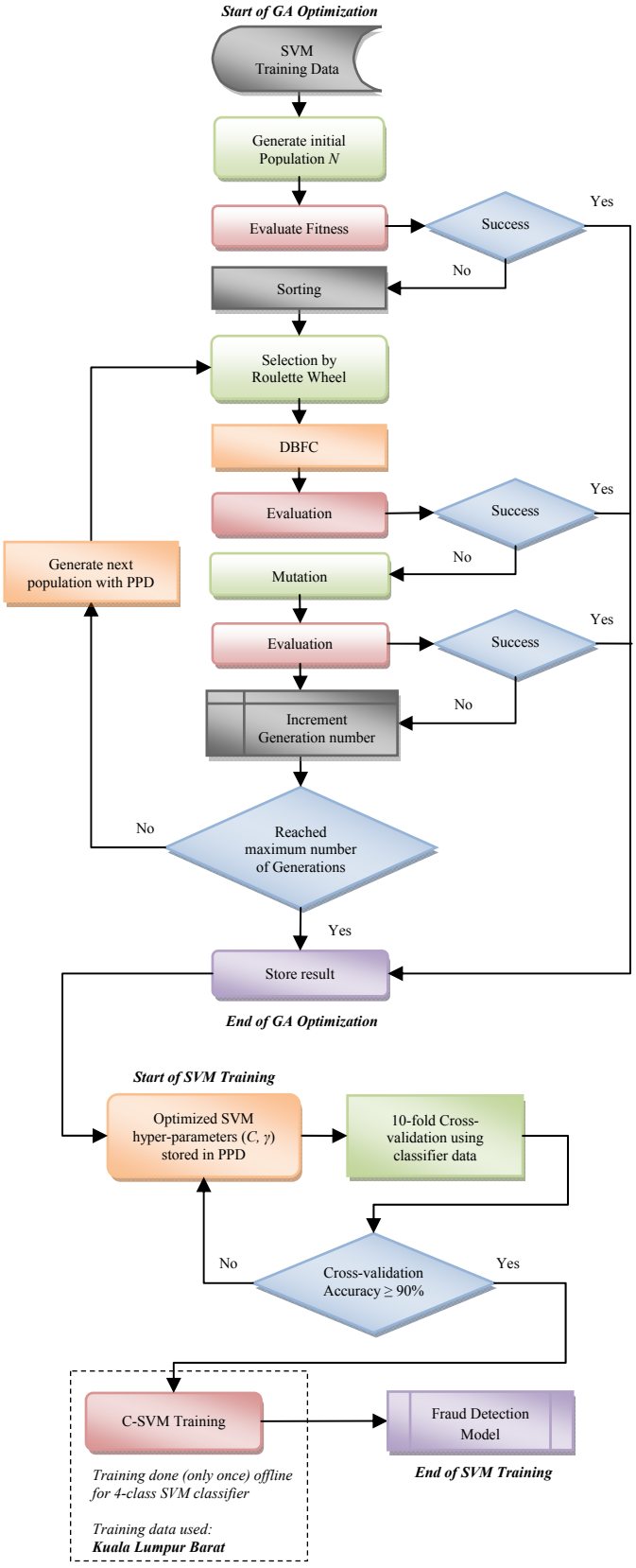

Fig. 6. Flowchart of proposed GA-SVM optimization and training engine

TABLE III
TEST BED USED FOR PREDICTING FRAUD ELECTRICITY CONSUMERS

City	Number of customers	Fraud cases previously detected by TNBD
Gua Musang	13,045	0
Kuala Krai	18,880	37
Kota Bharu	76,595	101

V. EXPERIMENTAL RESULTS

Pilot testing results obtained from TNBD for manual onsite inspection of NTL activities provided feedback that a hit-rate of 37% was achieved i.e., percentage of customers detected by TNBD as fraud from the total customers short listed. The other 63% cases met the following criteria:

- Replaced meters
- Abundant house
- Change of tenant
- Faulty meter wiring

Since load consumption patterns of these four (4) types of cases results in a similar load profiles as fraud cases, a logical decision based expert system was implemented to eliminate customers matching the four types of criteria. Inspection of load profiles for was performed manually to determine common characteristics distinguishing the four cases amongst all fraud cases detected. The expert system eliminated all four types of cases within short listed customers using conditions: $P_{class1} > 0.5$ and $P_{class2} > 0.26$ and $P_{class3} < 0.02$ and $P_{class4} < 0.025$ with load profile conditions listed in Table IV.

TABLE IV
EXPERT SYSTEM CONDITIONS FOR IMPROVING HIT-RATE

Condition	Description
$1.0 < X_{24} < 3.0$	X_{24} represents feature 24 i.e., 24th average daily kWh consumption feature value in the load profile.
$1.0 < X_{23} < 3.5$	X_{23} represents feature 23 i.e., 23rd average daily kWh consumption feature value in the load profile.
$X_{min} > 2$	X_{min} represents the minimum kWh consumption within the load profile.
$X_{max} > 15$	X_{max} represents the maximum kWh consumption within the load profile.
$X_{min\text{-}max} > 8$	$X_{min\text{-}max}$ represents the difference between maximum and minimum kWh consumptions of the load profile.

Performance of the fraud detection system by implementing the expert system increased the detection hit-rate from 37% to 62%. The increment resulted as a cause of eliminating the majority of unwanted customers. The GUI software developed for fraud detection is shown in Fig. 7.

VI. CONCLUSION

This paper presents a novel classification technique for detection of NTLs i.e., fraudulent electricity consumers. Results obtained show that the proposed GA-SVM framework can be used for reliable detection of fraudulent electricity consumers. The hybrid combination of GA-SVM provides a better solution for selecting optimal SVM hyper-parameters. With the implementation of the hybrid GA-SVM algorithm, the local optimum for finding the maximum Lagrangian was avoided. Furthermore, with the implementation of the proposed fraud detection system a hit-rate of 60% will be achievable. This will benefit TNB not only in improving its handling of NTLs, but will complement their existing on-going practices, and it is envisaged that tremendous savings will result from the use of the system.

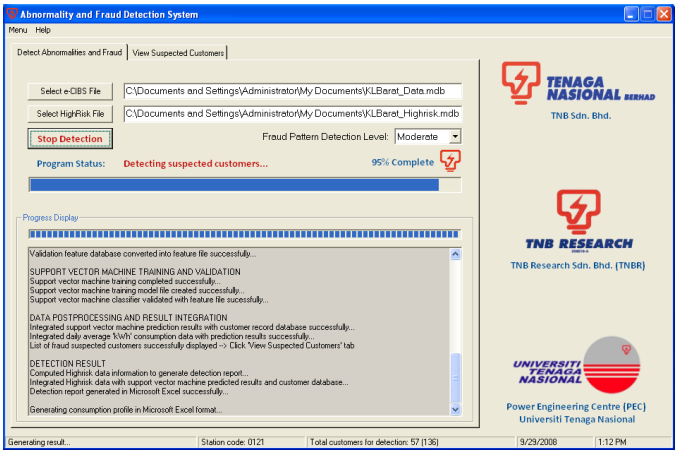

Fig. 7. GUI software developed fraud detection

REFERENCES

[1] R. Jiang, H. Tagaris, A. Lachsz, and M. Jeffrey, "*Wavelet Based Feature Extraction and Multiple Classifiers for Electricity Fraud Detection*" in Proc. of IEEE/PES Transmission and Distribution Conference and Exhibition 2002: Asia Pacific, Vol. 3, pp. 2251-2256.

[2] C. R. Paul, "System loss in a Metropolitan utility network" *IEEE Power Engineering Journal*, pp. 305-307, Sept. 1987.

[3] I. E. Davidson, A. Odubiyi, M. O. Kachienga, and B. Manhire, "Technical Loss Computation and Economic Dispatch Model in T&D Systems in a Deregulated ESI" *IEEE Power Eng. Journal*, Apr. 2002.

[4] A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power Utility Nontechnical Loss Analysis with Extreme Learning Machine Model" *IEEE Trans. on Power Systems*, Vol. 23, No. 3, pp. 946-955, Aug. 2008.

[5] J. W. Fourie and J. E. Calmeyer, "*A statistical method to minimize electrical energy losses in a local electricity distribution network*" in Proc. of the 7th IEEE AFRICON Conference Africa: Technology Innovation, Gaborone, Botswana, Sept. 2004.

[6] J. R. Filho, E. M. Gontijo, A. C. Delaiba, E. Mazina, J. E. Cabral, and J. O. P. Pinto, "*Fraud Identification in Electricity Company Customers Using Decision Trees*" in Proc. of 2004 IEEE International Conference on Systems, Man and Cybernetics, Vol. 4, pp. 3730-3734, Oct. 2004.

[7] J. R. Galvan, A. Elices, A. Munoz, T. Czernichow, and M. A. Sanz-Bobi, "*System for Detection of Abnormalities and Fraud in Customer Consumption*" in Proc. of the Electric Power Conference, Nov. 1998.

[8] A. H. Nizar, Z. Y. Dong, and J. H. Zhao, "*Load Profiling and Data Mining Techniques in Electricity Deregulated Market*" IEEE Power Engineering Society General Meeting, 2006, 18-22 Jun. 2006, pp. 1-7.

[9] D. Gerbec, S. Gasperic, I. Smon, and F. Gubina, "Allocation of the load profiles to consumers using probabilistic neural networks" *IEEE Trans. on Power Systems*, Vol. 20, No. 2, pp. 548-555, May 2005.

[10] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: Michigan Univ. Press, 1975 (Cambridge, MA: MIT Press, 1992).

[11] Y.-K. Kwon, B.-R. Moon and S.-D. Hong, "Critical heat flux function approximation using genetic algorithms" *IEEE Transactions on Nuclear Science*, Vol. 52, No. 2, pp. 535-545, Apr. 2005.

[12] V. Vapnik. *Statistical Learning Theory*, John Wiley & Sons, 1998.

[13] D. Wang, D. S. Yeung, E. C. C. Tsang, "Weighted Mahalanobis Distance Kernels for Support Vector Machines" *IEEE Transactions on Neural Networks*, Vol. 18, No. 5, pp. 1453-1462, Sept. 2007.

[14] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm